DOCUMENT RESUME

ED 331 855 TM 016 374

AUTHOR Schumacker, Randall E.; Harris, Mark J.
TITLE Reliability and Confidence Envelope Usage in Item Response Theory.
PUB DATE Apr 91
NOTE 21p.; Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, IL, April 3-7, 1991).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Ability; Equations (Mathematics); Item Banks; *Item Response Theory; *Mathematical Models; Test Construction; Testing Problems; *Test Reliability
IDENTIFIERS *Confidence Envelopes; Confidence Intervals (Statistics); *Item Characteristic Function

ABSTRACT

Designing a test using three-parameter item response theory (IRT) is discussed. A brief review of IRT is followed by a discussion of two types of test design: (1) selecting items using confidence envelopes (confidence envelope method); and (2) using item characteristic curves and their confidence intervals (test envelope method). The confidence envelope method and the test envelope method are evaluated based on their reliability coefficients, using a set of seven items. Results illustrate that the test envelope method, in which optimum ability levels are matched from an item bank, should result in a more reliable test. Six tables and three graphs illustrate the analysis. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED331855

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

RANDALL SCHUMACKER

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Reliability and Confidence Envelope Usage in Item Response Theory

Randall E. Schumacker
and
Mark J. Harris

University of North Texas

CORRESPONDENCE: Randall E. Schumacker, Ph.D.
Assistant Professor
Educational Foundations, Research,
and Special Education
P.O. Box 13857
University of North Texas
Denton, Texas 76203
(817) 565-3962

BITNET: FD44 @ UNTVM1

SPECIALTY: Measurement and Statistics

BEST COPY AVAILABLE

Paper Presented
at
American Educational Research Association
Chicago, Illinois
April 6, 1991

TM016374

ABSTRACT

The authors discuss designing a test using a recently developed approach in item response theory. A brief review of terms and formulae is followed by two types of test design. The first method suggests selecting items using confidence envelopes. The second method suggests using item characteristic curves and their confidence intervals. Using test reliability as the criteria, the second method is preferred for test design in item response theory.

Reliability and Confidence Envelope Usage

in

Item Response Theory

Item Response Theory

The three parameter IRT model is used to estimate P(θ) or the probability of a correct response to an item as follows:

$$P(\theta) = c + \frac{(1-c)}{1+e^{-1.7a(\theta-b)}}$$

The parameter c represents the probability that an examinee completely lacking in ability will answer the item correctly. It is a guessing parameter or pseudo chance score level. If an item cannot be answered correctly by guessing, then c = 0. The parameter b represents a location parameter. b is the position of the curve along the ability scale or item difficulty. The more difficult the item, the further the curve is to the right. The logistic curve has its inflection point at θ = b. When there is no guessing, b is the ability level where the probability of a correct answer is .5. When guessing occurs, b is the ability level where the probability of a correct answer is halfway between c and 1.00. The parameter a is proportional to the slope of the curve at the inflection point and equals 0.425 a (1 - c). It represents the discriminating power of the item or the degree to which the examinee response varies with ability level (Lord, 1980, p. 12-13).

The assumptions in item response theory are also of importance if estimating reliability. A reasonable assumption is that $P(\theta)$ increases as θ increases. Another suggests that an examinee's ability (θ) is all we need in order to determine the probability of success on a specific item. The assumption of local independence of items requires that any two items be uncorrelated when ability (θ) is fixed and follows directly from the assumption of unidimensionality for a test. Also, ability (θ) is probably not normally distributed for most groups of examinees, unidimensionality however, is a property of the items and does not cease to exist because the distribution of ability for a group changes (Lord, 1980, p. 19-20). Our concern is that the values of a, b, and c lie within a 95 % confidence interval.

A data set (Table 1) is used to illustrate our point (Wright and Stone, 1979, p. 31).

Insert Table 1 Here

An item analysis (Table 2) was conducted to estimate a, b, and c parameters (MicroCAT, 1986).

Insert Table 2 Here

Confidence Envelopes vs. Test Envelopes

A recently developed approach to test design in item response theory proposed using confidence envelopes (Thissen and Wainer, 1990). Accordingly, the authors state:

> "Confidence envelopes provide a description of the sampling variation of item response curves in the space of the fitted functions. They can be used to give the data analyst a clear idea of the class of item response curves that are compatible with the data. M-line plots may be used to show the width of the envelope, as well as the shapes and relative posterior density of the included curves." (p 126)

Each item characteristic curve is visually examined to see if it fits into the confidence envelope.

For example, given a = 45 degrees, b = 0 and c = .25 with difficulty ranging from + 1 θ to - 1 θ, the plot of the item characteristic curve can be examined to see if it lies within the upper and lower boundaries of the specific confidence envelope. The upper and lower boundaries can be computed using a 3-parameter IRT model (Lord and Pashley, 1988). Items for a completed test might appear as in Figure 1 and would include only those items selected within the upper and lower bounded confidence envelope.

Insert Figure 1 Here

Our approach uses the number of items or length of a test, L; the width of a test (Max θ - Min θ), W; and the average ability level of all examinees or height of a test, H. This approach models after known Rasch procedures (Wright and Stone, 1979). For the data set provided, Max θ = 3.803 and Min θ = -2.995, with W = 6.798 and H = 0. The optimum length is unknown.

A test envelope refers to the area in a plot of ability (θ) versus P(θ) bounded by the item characteristic curve of the lowest ability expected to the item characteristic curve of the highest ability expected (Figure 2).

Insert Figure 2 Here

The goal in test design would be to select item characteristic curves between the maximum and minimum ability (θ) such that the item and it's confidence interval cover the area without overlap (Note: Each item has it's own respective confidence interval). The item confidence interval can easily be computed using logistic regression (Hauck, 1983).

The authors derive the width of the confidence interval for a single item using a known three parameter IRT model procedure (Lord, 1980, pp. 66-67). We are however only interested in the width of the confidence interval (AB) at the point of inflection. For example, consider an item with a confidence interval around it at the point of inflection, b (Figure 3).

Insert Figure 3 Here

Derivation of AB

Given,

$$Tan\alpha = \frac{\overline{CB}}{\overline{AB}} = \frac{2(1.96)\sigma_{P(\theta)|\theta}}{\overline{AB}} = .425a(1-c),$$

we can use a and c to determine the slope of the line at the point of inflection, b (Lord, 1980).

Then,

$$\overline{AB} = 3.92 \frac{\sigma_{P(\theta)|\theta}}{.425a(1-c)}.$$

And since,

$$\sigma_{P(\theta)|\theta} = \frac{\sqrt{\sum_{\theta=1}^{N} P_i Q_i}}{N}.$$

Then,

$$\overline{AB} = \frac{3.92\sqrt{\sum_{\theta=1}^{N} P_i Q_i}}{.425a(1-c)N}$$

Therefore,

$$\theta_L = b - \frac{\overline{AB}}{2}$$

$$\theta_U = b + \frac{\overline{AB}}{2}.$$

The width of the confidence interval (the distance AB) describes the effectiveness of the test as a measure of ability (Lord, 1980, p. 66). The AB distances for 14 items in the data are in Table 3.

Insert Table 3 Here

The maximum AB distance is .452. The optimum length of the test would then be derived by W divided by maximum AB (6.798 divided by .452), or 16 items. The next step would involve computing the optimum b's for a test of H = 0, W = 6.798, L = 16. This can be accomplished by using the following formula (Wright and Stone, 1979, p. 140):

$$b_i = H + (W / 2) [(L - 2i + 1) / L]$$

The optimal item difficulties are in Table 4.

Insert Table 4 Here

Test construction requires an item pool and with these optimum difficulties included in the item bank information, item selection would be straightforward. Thus, a test envelope can be created with item characteristic curve information and item confidence intervals, respectively, using item response theory. This approach should make the application of item response theory in test construction easier for the practitioner.

Reliability of Methods

The confidence envelope method and the test envelope method are evaluated based upon their reliability coefficients. Lord's equation for reliability is used (Lord, 1980, p. 52):

$$\rho_{xx'} = \frac{\sum_{i=1}^{N}(\sum_{a=1}^{n} P_{ia})^2 - (\sum_{i=1}^{N}\sum_{a=1}^{n} P_{ia})^2/N}{\sum_{i=1}^{N}\sum_{a=1}^{n} P_{ia}Q_{ia} + \sum_{i=1}^{N}(\sum_{a=1}^{n} P_{ia})^2 - (\sum_{i=1}^{N}\sum_{a=1}^{n} P_{ia})^2/N}$$

The confidence envelope method, where all of the items were selected with the same value of b matched to the persons ability, is presented in Table 5. The test envelope method, where all of the items were selected with differing values of b for one person of ability (θ), is presented in Table 6.

Insert Table 5 Here

Insert Table 6 Here

If all of the items were selected to fit into the same confidence envelope for a = 1 and b = 0 for all seven items (Table 5), then:

$$r = \frac{12.25 - 1.75}{1.75 + 12.25 - 1.75} = .85714$$

If all the items were selected with certain confidence intervals to fit into a test envelope using optimum b values (Table 6), then:

$$r = \frac{12.25 - 1.75}{.58578 + 12.25 - 1.75} = .9471$$

Clearly, reliability increased because different non-overlapping items were selected to cover the range of ability measured (test envelope). The test envelope method, where the optimum ability levels are matched from an item bank, should result in a more reliable test.

References

Hauck, W. W. (1983). A note on confidence bands for the logistic response curve. The American Statistician, 37, 158-160.

Lord, F. M. & Pashley, P. J. (1988). Confidence bands for the three-parameter logistic item response curve (Research Report RR-88-67) Princeton, N. J.: Educational Testing Services.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. New Jersey: Lawrence Erlbaum Associates.

MicroCAT (1988). User's Manual for Iteman, Rascal, and Ascal. Assessment System Corporation: St. Paul, Minnesota.

Thissen, D. & Wainer, H. (1990). Confidence Envelopes for Item Response Theory. Journal of Educational Statistics, 15(2), 113-128.

Wright, B. D., & Stone, M. H. (1979). Best Test Design: Rasch Measurement. Mesa Press: Chicago, Il.

Table 1: 35 examinee responses to 18 items[a]

	Items																	
Person	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0
2	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0
3	1	1	1	1	1	1	1	1	1	0	0	1	0	0	0	0	0	0
4	1	1	1	1	0	0	1	0	1	0	0	0	0	0	0	0	0	0
5	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0
6	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0
7	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	0	0	0
8	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0
9	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0
10	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0
11	1	1	1	0	1	1	1	1	1	0	0	0	0	0	0	0	0	0
12	1	1	1	1	1	0	1	0	1	1	0	0	0	0	0	0	0	0
13	1	1	1	1	1	0	0	1	1	1	1	1	1	0	0	0	0	0
14	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0
15	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0
16	1	1	1	1	1	1	1	1	1	0	1	0	0	0	0	0	0	0
17	1	1	1	1	0	1	1	1	1	1	0	0	0	0	0	0	0	0
18	1	1	1	1	1	1	1	1	1	1	0	0	1	0	0	0	0	0
19	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0
20	1	1	1	1	1	1	1	1	1	1	0	0	1	0	0	0	0	0
21	1	1	1	1	1	1	1	1	1	1	1	0	1	0	0	0	0	0
22	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0
23	1	1	1	1	1	1	1	1	1	1	0	0	1	1	0	0	0	0
24	1	1	1	1	1	1	1	1	1	1	1	0	1	0	0	1	1	0
25	1	1	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0
26	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0
27	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0
28	1	1	1	1	1	1	1	1	1	0	1	0	0	0	0	0	0	0
29	1	1	1	1	1	1	0	0	1	1	1	0	0	1	0	0	0	0
30	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0
31	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0
32	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0
33	1	1	1	1	0	0	1	0	0	1	0	0	0	0	0	0	0	0
34	1	1	1	1	1	1	1	1	1	1	0	1	0	1	0	0	0	0
35	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

[a] (Wright and Stone, 1979, p. 31)

Table 2: IRT parameter estimates for data[a]

Item	a	b	c
1	item deleted		
2	item deleted		
3	item deleted		
4	1.547	-1.573	0.240
5	1.475	-1.298	0.250
6	1.740	-1.074	0.250
7	1.578	-1.336	0.240
8	1.938	-0.818	0.180
9	1.787	-1.189	0.220
10	1.801	-0.161	0.260
11	1.943	0.757	0.190
12	2.500	3.000	0.190
13	2.500	3.000	0.200
14	1.567	3.000	0.130
15	1.622	2.289	0.030
16	1.622	2.289	0.030
17	1.622	2.289	0.030
18	item deleted		

[a] (MicroCAT, 1988)

Table 3: Width ($\overline{AB}$) of the confidence interval at the point of inflection b

Item	Width ($\overline{AB}$)
4	.288
5	.367
6	.329
7	.324
8	.292
9	.290
10	.452
11	.420
12	.302
13	.312
14	.389
15	.204
16	.204
17	.204

Table 4: Optimum b_i[a]

Item	b_i
1	3.19
2	2.76
3	2.34
4	1.91
5	1.49
6	1.06
7	0.64
8	0.21
9	-0.21
10	-0.64
11	-1.06
12	-1.49
13	-1.91
14	-2.34
15	-2.76
16	-3.19

[a] $\underline{t}(\underline{h}=0, \underline{w}=6.798, \underline{l}=16)$

Table 5: Confidence envelope method[a]

b_i	$P(\theta)$	$Q(\theta)$	$P(\theta)Q(\theta)$
0	.50	.50	.25
0	.50	.50	.25
0	.50	.50	.25
0	.50	.50	.25
0	.50	.50	.25
0	.50	.50	.25
0	.50	.50	.25
Sum			1.75

[a] Assumes equal item ability, discrimination, and difficulty

Table 6: Test Envelope Method[a]

b_i	$P(\theta)$	$Q(\theta)$	$P(\theta)Q(\theta)$
-3	.99394	.00606	.00602
-2	.96770	.03230	.03126
-1	.84553	.15447	.13061
0	.50000	.50000	.25000
1	.15447	.84553	.13061
2	.03230	.96770	.03126
3	.00606	.99394	.00602
Sum			.58578

[a] Assumes equal item ability and discrimination

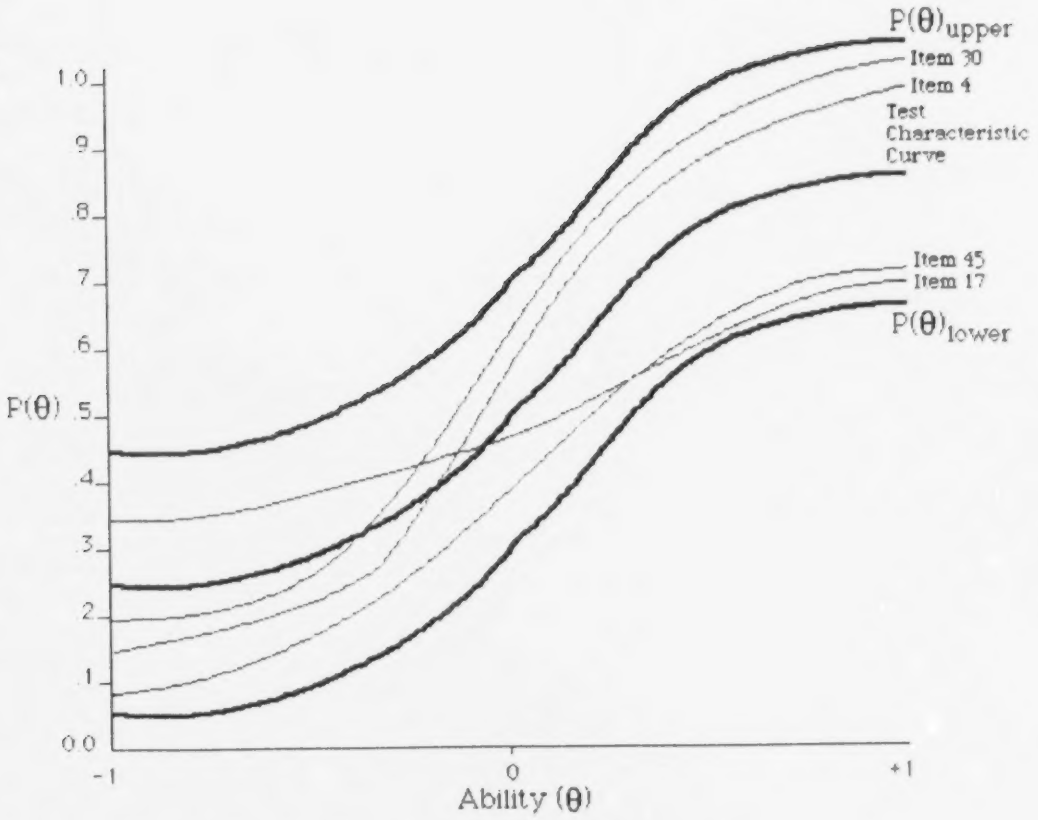

Figure 1: Confidence Envelope

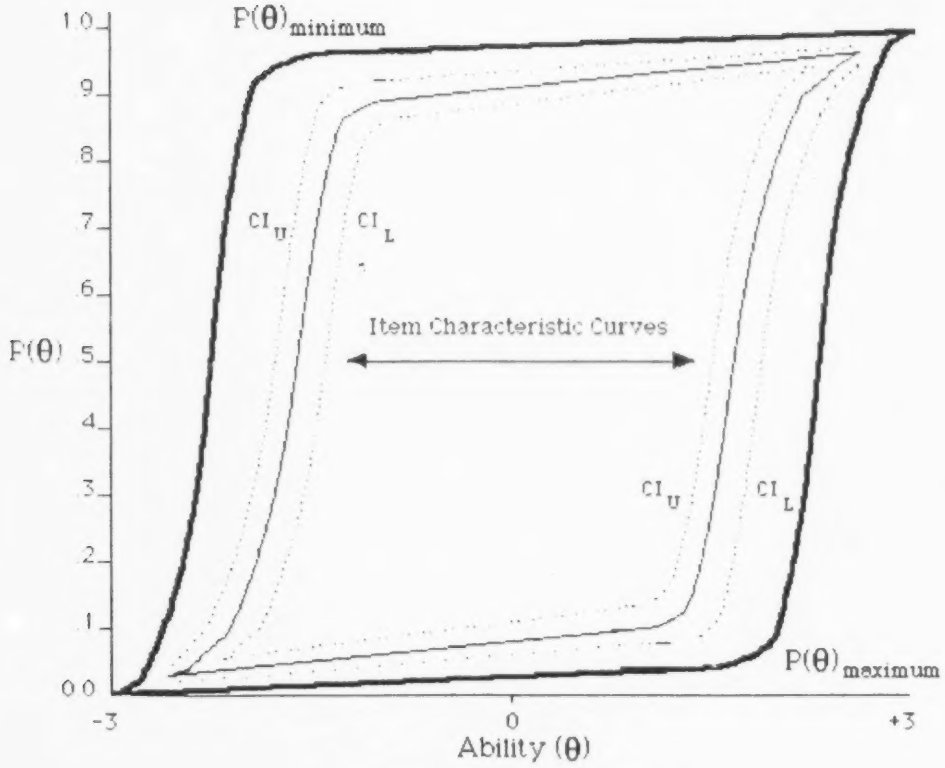

Figure 2. Test Envelope

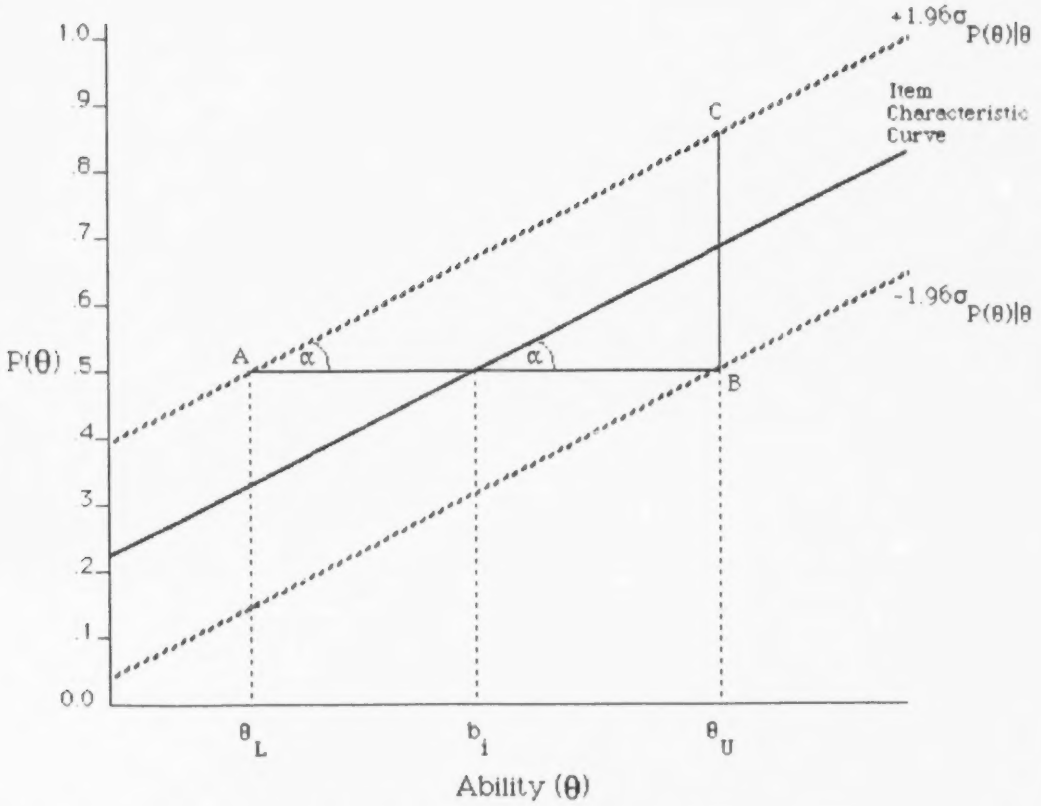

Figure 3: Confidence interval width for a single item at point of inflection b_i